Détection de la négation : corpus français et apprentissage supervisé

Clément Dalloux*, Natalia Grabar**, Vincent Claveau*

*IRISA - CNRS, Campus de Beaulieu, 35042 Rennes, France,
prenom.nom@irisa.fr,
http://people.irisa.fr/Prenom.Nom/
**UMR 8163 STL CNRS, Université de Lille 3 - France,
natalia.grabar@univ-lille3.fr,
http://natalia.grabar.free.fr

Résumé. La détection automatique de la négation fait souvent partie des prérequis dans les systèmes d'extraction d'information, notamment dans le domaine biomédical. Cet article présente deux contributions liées à ce problème. Nous présentons d'une part un corpus constitué d'extraits des protocoles d'essais cliniques en français, dédié aux critères d'inclusion de patients. Les marqueurs de négation et leurs portées y ont été annotés manuellement. Nous présentons d'autre part une approche neuronale supervisée pour extraire ces informations automatiquement. Cette approche est validée en l'appliquant à des données de l'état de l'art en anglais sur lesquelles elle montre de très bons résultats ; appliquée sur nos données en français, elle obtient des performances comparables.

1 Introduction

Les marqueurs de la négation sont d'habitude constitués d'un ou de plusieurs mots qui souvent modifient la polarité et donc le sens de la phrase. Notons que la négation peut également être marquée par les préfixes. Dans les exemples qui suivent les marqueurs sont soulignés. Même si la détection des marqueurs est une tâche assez complexe, dû entre autres à leur variété et ambiguïté, elle n'est pas suffisante en soi. Ainsi, en plus de la détection des marqueurs, il est également nécessaire de calculer leur portée : de décider quel est l'effet du marqueur et si cet effet s'étend sur toute la phrase ou sur une partie de cette phrase. Dans les exemples qui suivent, la portée est marquée entre les crochets. Pour ces diverses raisons, la détection des informations liées à la négation présente de multiples difficultés, comme par exemple :

— la négation peut être annulée par un adverbe de fréquence (exemple en (1)),
— les préfixes *an-, in-, im-, ir-, dis-*, etc peuvent également marquer la négation (exemples en (4-5)),
— les marqueurs peuvent fonctionner sous différents régimes et leur portée peut aller à droite ou à gauche, s'étendre des deux côtés (exemple en (2)), être discontinue (exemple en (3)) ou se chevaucher (exemples en (4-5)).

1. *Une discipline pas toujours suivie par les personnes concernées.*

2. *Asthme : [les hormones protègent] les hommes, pas [les femmes].*
3. *L'étude vise également à vérifier que [cette information] est recevable, ne [génère] pas [de stress],[...].*
4. *Le traitement par tazemetostat continuera jusqu'à progression de la maladie ou l'apparition d'[un effet] in[désirable] inacceptable.*
5. *Le traitement par tazemetostat continuera jusqu'à progression de la maladie ou l'apparition d'[un effet indésirable] in[acceptable].*

Dans le domaine biomédical en particulier, la négation joue un rôle important. Dans le cas d'essais cliniques par exemple, elle peut fournir un critère déterminant pour recruter ou non un patient. Nous proposons donc de travailler sur cette problématique. Dans ce qui suit, nous présentons d'abord les travaux qui existent et décrivons les corpus exploités. Nous décrivons ensuite les méthodes proposées et discutons les résultats obtenus. Nous terminons avec une conclusion et les pistes pour le travail futur.

2 Travaux existants

Dans Dalloux (2017), nous avons exploré les différents corpus et approches dédiés à la détection des marqueurs ainsi que de la portée de la négation. Dans les sections suivantes, nous revenons brièvement sur ces travaux.

2.1 Les données

Ces dernières années, avec la démocratisation des techniques d'apprentissage supervisé, plusieurs corpus de spécialité ont été annotés afin d'entraîner des modèles pour la détection automatique de la négation en anglais. Les corpus se divisent en deux catégories : les corpus avec marqueurs et portées annotées, comme Bioscope (Vincze et al., 2008) et *SEM-2012 par exemple, ainsi que les corpus se focalisant sur le contexte entourant les entités nommées tels que Shapn, i2b2 et mipacq.

2.2 Détection automatique

D'une manière générale, il existe deux familles d'approches pour aborder la détection automatique de la négation. L'utilisation de systèmes experts, *NegEx* initiée par Chapman et al. (2001) et *Negfinder* proposé par Mutalik et al. (2001). Plus récemment, de nombreux travaux utilisent la classification par apprentissage supervisé à l'aide de méthodes telles que les champs aléatoires conditionnels (*Conditional Random Fields* ou CRFs), les machines à vecteurs support (SVM) ou les réseaux de neurones (Velldal et al., 2012; Read et al., 2012; Packard et al., 2014; Fancellu et al., 2016).

3 Description et préparation du corpus

Notre corpus contient des protocoles d'essais cliniques en français. Ces protocoles ont été récupérés grâce au registre des essais cliniques de l'hôpital Gustave Roussy [1], ainsi que l'Insti-

1. https ://www.gustaveroussy.fr/fr/essais-cliniques

Forme	Lemme	POS Tag	Marqueur	Portée
Dans	dans	PRP	_	_
les	le	DET :ART	_	_
formes	forme	NOM	_	_
peu	peu	ADV	_	_
sévères	sévère	ADJ	_	_
,	,	PUN	_	_
une	un	DET :ART	_	une
antibiothérapie	antibiothérapie	NOM	_	antibiothérapie
spécifique	spécifique	ADJ	_	spécifique
n’	ne	ADV	n’	_
est	être	VER :pres	_	est
habituellement	habituellement	ADV	_	habituellement
pas	pas	ADV	pas	_
nécessaire	nécessaire	ADJ	_	nécessaire
.	.	SENT	_	_

TAB. 1 – *Extrait du corpus, multiples colonnes à l’instar de *SEM-2012*

tut National du Cancer [2]. Nous avons décidé de nous concentrer sur les critères d’inclusion des patients et sur la description du déroulement des essais. Le corpus contient, à ce jour, 4 198 phrases, dont 650 négatives (15,48 %) issues de plusieurs centaines de documents/pages web. Le corpus est pré-traité avec TreeTagger (Schmid, 1994) pour effectuer l’étiquetage morpho-syntaxique et la lemmatisation. Le corpus a été annoté manuellement pour marquer les informations liées à la négation : les marqueurs et leur portée.

Dans le tableau 1, nous présentons un exemple de phrases avec différents niveaux d’informations linguistiques (la forme, le lemme, l’étiquette morpho-syntaxique ou le POS-tag) et les catégories de l’annotation manuelle (le marqueur de la négation et sa portée).

4 Méthodes

Nous abordons la tâche de détection de la portée de la négation comme une tâche de classification : les informations recherchées doivent être prédites grâce à l’algorithme d’apprentissage supervisé, que nous décrivons dans la suite de cette section. Nous présentons ensuite notre approche.

4.1 Réseau de neurones récurrents

Un réseau de neurones récurrents (RNN) est motivé par le fait qu’un être humain raisonne en s’appuyant sur les connaissances qu’il a acquises et qui restent dans la mémoire. Si les réseaux de neurones « classiques » ne sont pas capables de copier ce processus, les RNN répondent à ce problème. En effet, ce sont des réseaux qui bouclent, ce qui permet aux informations de persister.

Parmi les RNN, les réseaux *Long short-term memory* (LSTM), proposé par Hochreiter et Schmidhuber (1997), sont les plus efficaces pour apprendre des dépendances de long terme et sont donc plus à même de résoudre, par exemple, le problème de la portée discontinue.

2. http ://bit.ly/2ypwAoA

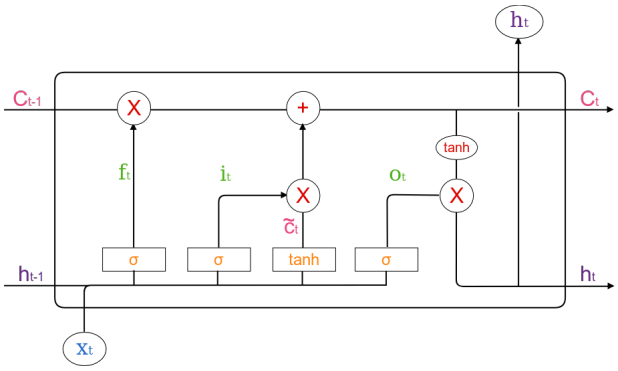

FIG. 1 – *Cellule LSTM.*

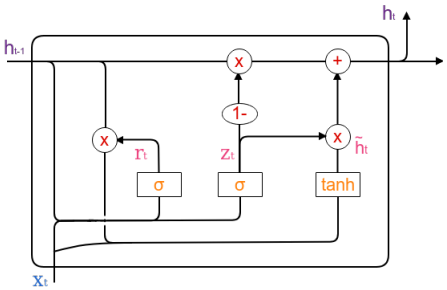

FIG. 2 – *Cellule GRU.*

Par ailleurs, les cellules LSTM sont plus efficaces pour retenir les informations utiles lors de la rétropropagation du gradient, qui permet de corriger les différences entre les prédictions sortantes et celles désirées en calculant le gradient de l'erreur pour chaque neurone, de la dernière couche vers la première. La figure 1 illustre les quatre couches interactives (sigmoïde et tanh), les trois portes (oubli f, entrée i et sortie o) et les opérations *pointwise* qui traitent le vecteur x à l'intérieur d'une cellule LSTM à un temps t.

Proposé par Cho et al. (2014), le réseau Gated Recurrent Unit (GRU) est une variante du LSTM où les portes d'oubli et d'entrée sont fusionnées en une unique porte de mise à jour. L'état de la cellule C ainsi que l'état caché h sont aussi fusionnés (voir figure 2). Le modèle produit est plus simple que celui d'un LSTM standard. En pratique, cette approche permet de réduire le temps de calcul d'un modèle tout en conservant des résultats équivalents à ceux d'un réseau LSTM.

4.2 Notre approche

Notre approche est basée sur l'utilisation des réseaux de neurones récurrents présentés précédemment. En outre, nos RNN sont bidirectionnels, c'est-à-dire opérant dans le sens de

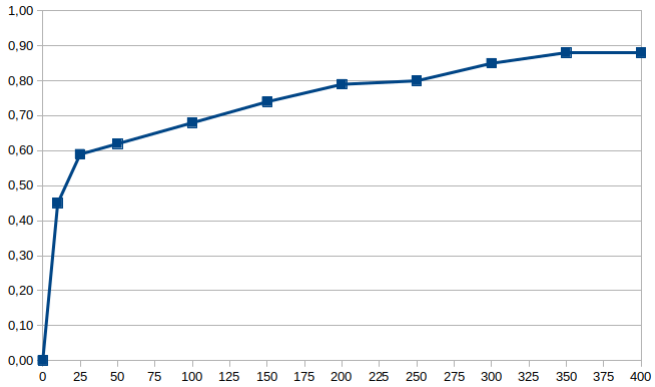

FIG. 3 – *Progression de la F-mesure selon le nombre d'exemples lors de l'entraînement*

lecture et dans le sens contraire. La passe arrière est particulièrement importante dans le cas de la détection de la portée puisque les mots affectés peuvent se trouver avant le marqueur.

Implémenté à l'aide de *Tensorflow* (Abadi et al. (2016)), notre système comprend une version adaptée au français du LSTM bidirectionnel de Fancellu et al. (2016), ainsi que son adaptation en GRU bidirectionnel. En sortie, la prédiction est assurée soit par une couche *softmax*, qui est la méthode la plus courante, soit par une couche CRF, une méthode qui semble être particulièrement efficace pour l'étiquetage de séquences.

Nos réseaux s'entraînent uniquement sur les phrases porteuses d'une négation. Le système de base prend en entrée une instance $I(n, c, t)$ où chaque mot est représenté par un vecteur n (*word-embedding*), un vecteur c qui détermine si le mot fait partie d'un marqueur (*cue-embedding*), ainsi que d'un vecteur t qui est la représentation vectorielle de l'étiquetage morpho-syntaxique pour chaque mot (*postag-embedding*).

Pour chaque système, nous utilisons les mêmes paramètres d'entraînement définis de façon empirique. Nos embeddings sont de dimension $k = 50$. La couche cachée compte 200 unités (400 pour le BiLSTM qui nécessite deux couches cachées concaténées). 30 périodes d'entraînement permettent d'atteindre le meilleur score F_1 possible sur l'ensemble de validation.

Le corpus annoté est segmenté en trois parties définies empiriquement : un ensemble d'entraînement (350 phrases), un ensemble de validation (100 phrases) et un ensemble de test (200 phrases). Comme l'illustre la figure 3, dans l'état actuel du corpus, il apparaît que l'augmentation de la quantité d'exemples de l'ensemble d'entraînement n'apporte plus de bénéfices au-delà de 350 phrases.

Les résultats obtenus sur le corpus de test sont évalués contre les données de référence avec les mesures d'évaluation classiques : la précision P, qui quantifie la pertinence de l'étiquetage, le rappel R, qui quantifie la sensibilité de l'étiquetage, ainsi que la moyenne harmonique de la précision et du rappel noté F_1. Nous indiquons aussi l'écart type des moyennes de F-mesures de nos *runs*.

	Mots étiquetés				Portées exactes			
Système	P	R	F_1	s_{F_1}	P	R	F_1	s_{F_1}
BiLSTM	89,07	86,54	**87,68**	1,56	100	60,10	74,93	5,37
BiLSTM+CRF	87,69	**87,20**	87,40	1,24	100	**61,95**	**76,48**	2,37
BiGRU	89,69	82,68	85,95	1,26	100	58,42	73,74	1,30
BiGRU+CRF	**90,93**	82,83	86,67	1,08	100	59,76	74,77	2,79

TAB. 2 – *Résultats des différentes approches sur notre corpus (moyenne de 3 runs sur trois ensembles de données aléatoires). Écart type des moyennes de F-mesures s_{F_1}. Les résultats sont donnés en pourcentage et les meilleurs scores sont indiqués en gras.*

	Mots étiquetés			Portées exactes		
Système	P	R	F_1	P	R	F_1
Read et al. (2012)	81,99	88,81	85,26	87,43	61,45	72,17
Lapponi et al. (2012)	86,03	81,55	83,73	85,71	62,65	72,39
Packard et al. (2014)	86,1	**90,4**	88,2	98,8	**65,5**	**78,7**
Fancellu et al. (2016)	**92,62**	85,13	88,72	**99,40**	63,87	77,7
BiLSTM+CRF (nous)	91,24	87,10	**89,12**	**100**	62,5	76,92

TAB. 3 – *Comparaison des systèmes les plus performants sur les données de test de *SEM-2012. Les résultats sont donnés en pourcentage et les meilleurs scores sont indiqués en gras.*

5 Résultats et discussion

Le tableau 2 présente les résultats obtenus avec notre approche, tandis que le tableau 3 présente les résultats obtenus par les travaux existants effectués sur les données de **SEM-2012*.

Premièrement, comme nous nous y sommes attendus, les LSTM donnent de meilleurs résultats que les GRU. En effet, la revue des performances des RNN sur de multiples tâches faite par Jozefowicz et al. (2015) rapporte que les GRU donnent toujours de meilleurs résultats que les LSTM, excepté pour la tâche de modélisation du langage. Bien que toutes les portes aient un impact positif sur les résultats, les expérimentations proposées dans ce travail montrent que c'est la porte d'oubli qui donne l'avantage au LSTM. Par ailleurs, la prédiction par CRF donne, pour le LSTM comme le GRU, de meilleurs résultats que la couche *softmax* classique dans la détection de portées exactes. Les CRF sont particulièrement efficaces pour l'étiquetage de séquences, ce qui pourrait expliquer le léger gain (1/1,5 points) obtenu.

Concernant l'ensemble de données de **SEM-2012*, nous obtenons un score F_1 légèrement supérieur à celui du BiLSTM de Fancellu et al. (2016) en terme de mots correctement étiquetés. Cela s'explique par un meilleur équilibre entre précision et rappel. Cela dit, notre score est légèrement inférieur en terme de portées exactes.

Un examen des résultats permet d'isoler des cas récurrents d'erreurs. L'exemple ci-dessous, dans lequel le sujet est séparé par plusieurs mots de la négation le concernant, impacte le rappel.

— GOLD : *L'objectif de cet essai est de comparer l'efficacité et la tolérance de la gemcitabine administrée seule ou en association avec de l'oxaliplatine chez [des patients] ayant un cancer de la vessie à un stade avancé et ne [pouvant être traité par du cisplatine].*

— PRED : *L'objectif de cet essai est de comparer l'efficacité et la tolérance de la gemcitabine administrée seule ou en association avec de l'oxaliplatine chez des patients ayant un cancer de la vessie à un stade avancé et ne [pouvant être traité par] du [cisplatine].*

Un deuxième exemple illustre un autre problème d'erreur impactant la précision. Ici, le modèle prédit que *« et »* fait porter la négation sur l'énoncé qui le suit car il a sans doute rencontré ce type de phrase lors de son entraînement.

— GOLD : *[La prise en charge thérapeutique] ne [sera] pas [modifiée par l'étude] et sera conforme aux référentiels.*
— PRED : *[La prise en charge thérapeutique] ne [sera] pas [modifiée par l'étude et] sera [conforme aux référentiels].*

6 Conclusions & perspectives

Les travaux sur la détection automatique de la négation en langue anglaise par apprentissage supervisé se sont multipliés ces dernières années. Dans cet article, après avoir présenté les difficultés liées à cette tâche et après un bref rappel des travaux existants, nous avons présenté un nouveau corpus de données biomédicales, de langue française, annotées avec les informations sur la négation (les marqueurs et leur portée). Avant sa diffusion auprès de la communauté de la recherche, le corpus sera finalisé grâce à l'intégration de données nouvelles provenant de protocoles d'essais cliniques et possiblement d'autres sources, suivi d'une étape d'harmonisation des annotations.

Une autre contribution de notre travail consiste en exploitation des réseaux de neurones récurrents pour la reconnaissance automatique des marqueurs et de leur portée. Les expériences sont effectuées et évaluées dans deux langues : en anglais, qui dispose de données de référence indépendantes, et en français, qui n'a pas connu beaucoup de travaux de ce type. Notre système montre des performances très proches du meilleur système de l'état de l'art en anglais, et des performances très proches en français. Nos travaux indiquent également que les architectures neuronales fondées sur les LSTM sont plus efficaces que les GRU dans la tâche de détection de la portée. En outre, la couche de sortie CRF semble apporter de légers gains en terme de rappel par rapport à la couche *softmax* grâce à une meilleure prise en compte des dépendances entre labels.

À ce jour, aucun système expert, tel que Deléger et Grouin (2012), n'a été testé sur nos données. C'est une lacune qu'il nous faudra combler avant la finalisation du corpus. Par ailleurs, une première version de notre système est désormais utilisable via *ALLGO* : `https://allgo.inria.fr/webapps/173`

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Références

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et et al. (2016). Tensorflow : Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv :1603.04467*.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, et B. G. Buchanan (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics 34*(5).

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et Y. Bengio (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*.

Dalloux, C. (2017). Détection de l'incertitude et de la négation : un état de l'art. In *19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, pp. 94–107.

Deléger, L. et C. Grouin (2012). Detecting negation of medical problems in french clinical notes. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*.

Fancellu, F., A. Lopez, et B. Webber (2016). Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Volume 1.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural Comput. 9*(8).

Jozefowicz, R., W. Zaremba, et I. Sutskever (2015). An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2342–2350.

Lapponi, E., E. Velldal, L. Øvrelid, et J. Read (2012). *Uio 2 : sequence-labeling negation using dependency features*. Association for Computational Linguistics.

Mutalik, P. G., A. Deshpande, et P. M. Nadkarni (2001). Use of general-purpose negation detection to augment concept indexing of medical documents : a quantitative study using the umls. *Journal of the American Medical Informatics Association : JAMIA 8*(6).

Packard, W., E. M. Bender, J. Read, S. Oepen, et R. Dridan (2014). *Simple Negation Scope Resolution through Deep Parsing : A Semantic Solution to a Semantic Problem.*

Read, J., E. Velldal, L. Øvrelid, et S. Oepen (2012). *Uio 1 : Constituent-based discriminative ranking for negation resolution*. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-ofispeech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.*

Velldal, E., L. Øvrelid, J. Read, et S. Oepen (2012). Speculation and negation : Rules, rankers, and the role of syntax. *Computational Linguistics 38*(2).

Vincze, V., G. Szarvas, R. Farkas, G. Móra, et J. Csirik (2008). The bioscope corpus : biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics 9*.

Summary

Automatic detection of negated content is often a pre-requisite in information extraction systems, especially in the biomedical domain. This paper proposes two main contributions in this field. It first introduces a corpus built with excerpts from clinical trial protocols in French, describing the inclusion criteria of patients. The corpus is manually annotated for marking up the negation cues and their scope. Secondly, a supervised neural approach is proposed to acquire the negations. This approach is validated on English data on which it outperforms existing approaches; when applied on our French data, it also yields very good results.